EVALUATION OF A SPOKEN DIALOGUE SYSTEM FOR VIRTUAL REALITY CALL FOR FIRE TRAINING

Susan M. Robinson, Antonio Roque, Ashish Vaswani
David Traum, Charles Hernandez, Bill Millspaugh

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **01 NOV 2006**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE
Evaluation Of A Spoken Dialogue System For Virtual Reality Call For Fire Training

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Creative Technologies, University of Souhern California
13274 Fiji Way, Marina del Rey, CA, 90292

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM002075., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**30**	

Outline

- Virtual Reality Call for Fire Training
- The Radiobot-CFF System
- Evaluation method
- Evaluation Results
- Next Steps

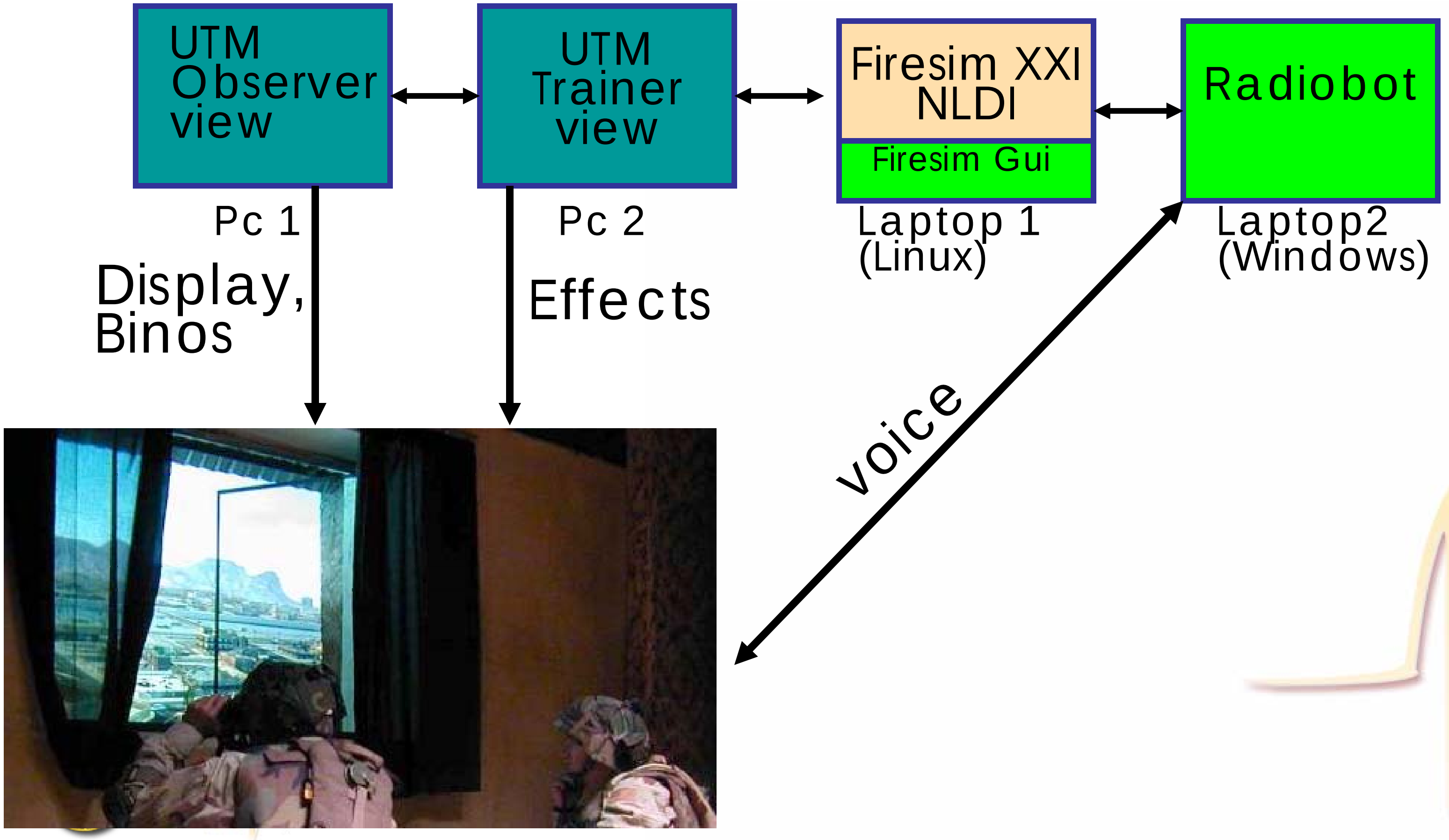
Radiobots: Project History

- 2004: Piloted within ICT Mission Rehearsal Exercise (MRE) Project
  - Simple dialogue systems for radio characters
  - Output through radio
- 2004-2005: seedling effort
  - Further development of MRE radiobots
  - Analysis of radiobot domains & tools
    - Focus on call for fire
  - Tools for data collection & semi-automatic operation
  - Initial data collection at Ft Sill and analysis
- 2005 - 2006: Radiobots for JFETS: Radiobot-CFF

Radiobots for JFETS: Team members

- USC ICT (Dr. David Traum, Antonio Roque, Susan Robinson, Dr Anton Leuski, Jarrell Pair, Tae Yoon, Dr Bilyana Martinovski, Ashish Vaswani, Sudeep Gandhe, Emily Flores, Jillian Gerten)
  - overall integration & management
  - dialogue systems
  - corpus creation & development
  - evaluation
- USC SAIL (Dr. Shri Narayanan, Vivek Sridhar, Shankar Anathakrishnan)
  - speech processing
- TechMasters Inc (TMI) (Bill Millspaugh)
  - FireSIM XXI simulation
  - Text to tactical messaging (NLDI)
- ARL-HRED (Charles Hernandez, Dr Janet Sutton)
  - Evaluation
- With help from Ft Sill Battle Lab & Techrizon

System Architecture: Hardware and User Interaction



UTM Observer view

Pc 1

UTM Trainer view

Pc 2

Firesim XXI NLDI

Firesim Gui

Laptop 1 (Linux)

Radiobot

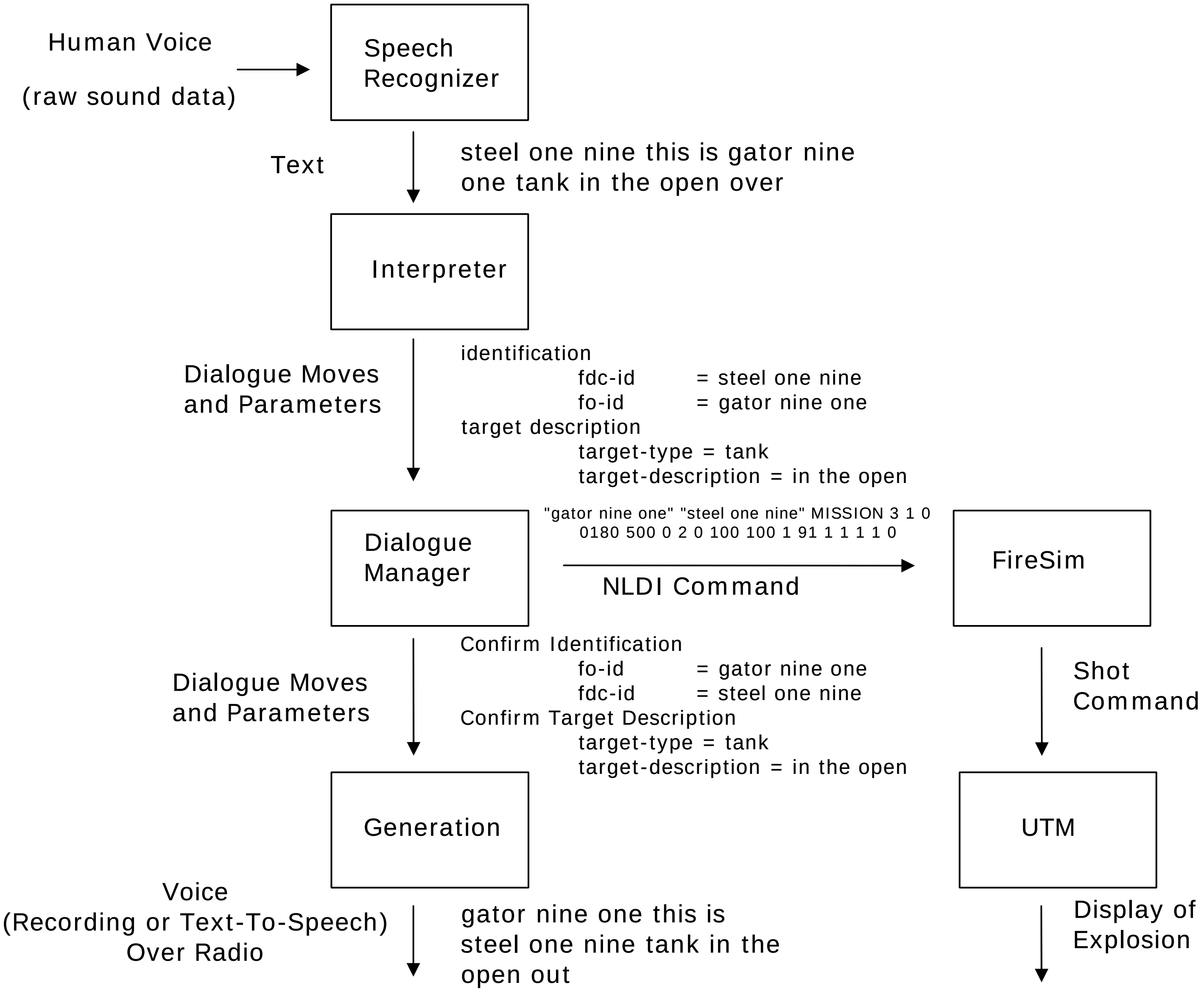
Laptop2 (Windows)

Display, Binos

Effects

voice

System Architecture:
Software components and dataflow



Human Voice

(raw sound data)

Speech Recognizer

Text

steel one nine this is gator nine one tank in the open over

Interpreter

Dialogue Moves and Parameters

identification
  fdc-id = steel one nine
  fo-id = gator nine one
target description
  target-type = tank
  target-description = in the open

Dialogue Manager

"gator nine one" "steel one nine" MISSION 3 1 0 0180 500 0 2 0 100 100 1 91 1 1 1 1 0

NLDI Command

FireSim

Shot Command

UTM

Display of Explosion

Dialogue Moves and Parameters

Confirm Identification
  fo-id = gator nine one
  fdc-id = steel one nine
Confirm Target Description
  target-type = tank
  target-description = in the open

Generation

Voice
(Recording or Text-To-Speech)
Over Radio

gator nine one this is steel one nine tank in the open out

Example Radiobot Interactions

G91: steel one niner this is gator niner one , adjust fire over ,

S19: gator nine one this is steel one nine , adjust fire out ,

G91: grid four five one , three six four over

S19: grid four five one three six four out ,

G91: one z_s_u in the open , i_c_m in effect over ,

S19: one z_s_u in the open , i_c_m in effect out .

S19: message to observer . kilo alpha high explosive four rounds . adjust fire target number alpha bravo one zero zero zero over ,

G91: message to observer , kilo alpha , high explosive in effect four rounds , target number alpha bravo one zero zero break ,

S19: shot over ,

G91: shot out ,

S19: splash over ,

G91: splash out

G91: steel one nine this is gator nine one , adjust fire polar over ,

S19: gator nine one this is steel one nine , adjust fire polar out ,

G91: direction five nine seven zero , distance four eight zero over ,

S19: direction five nine seven zero , distance four eight zero out ,

G91: one b_m_p in the open , d_p_i_c_m in effect over .

S19: one b_m_p in the open . i_c_m in effect out .

S19: message to observer . kilo bravo high explosive four rounds . adjust fire target number alpha bravo one zero zero two over

G91: message to observer , kilo alpha quick in effect h_e four rounds , target number alpha bravo one thousand two over ,

S19: shot target number alpha bravo one zero zero two over ,

G91: shot out ,

Evaluation Goals

- Measures of performance of system and components
- Measures of effectiveness of system for use in training in the JFETS Urban Terrain Module
- Measures of User Satisfaction
- Identify areas of needed improvement

Evaluation Metrics

- System Performance Metrics
  - mission completion, timing to fire, accuracy, transmission quality
- Component Performance Metrics
  - ASR, interpreter, dialogue manager, generator
- Subjective Data
  - Questionnaires

Evaluation Conditions

- Automated: radiobot as FSO, automatically sends mission information to Firesim
- Semi-automated: As above, but fills in form for human operator to review (possibly correct) and submit
- Human control: Human FSO engages in radio dialogues and human operator sends missions through Firesim

Evaluation Sessions

- Preliminary Evaluation Nov 2005
  - 34 students in UTM training
  - Focused on semi-auto condition and refining user questionnaire

- Final Evaluation Jan-Feb 2006
  - 29 volunteers from Ft Sill, some repeat subjects across conditions
  - Demographic and user surveys for each session
  - 2 subjects per group, FO and RTO each did 2 missions then switched roles.
  - Conditions were varied across groups

Evaluation Data Overview

- Eval 1: Jan 2006
  - 20 sessions (10 teams)
  - 4 human, 8 semi-auto, 8 auto
- Eval 2: Feb 2006
  - 27 sessions (14 teams)
  - 6 human, 9 semi-auto, 12 auto

Evaluation Results: Mission Performance

- Average time to fire:

  Human: 1 min 46

  Semi: 2 min 19

  Auto: 1 min 44

- Task completion rate:

  Human: 100%

  Semi: 98%

  Auto: 86%

- Accuracy rate:
  - Human: 100%
  - Semi: 97%
  - Auto: 92%

Transmission Quality

Session	System transmissions	Acks req	% Acks	Repair Requests	Correct responses	Flawless Responses	Flawless transmissions
W1-2	27	12	100%	8%	92%	58%	82%
W3-1	26	14	100%	14%	93%	50%	73%
T2-2	15	8	88%	0	71%	71%	87%
T4-2	21	13	85%	0	91%	46%	71%
T5-2	67	39	97%	11%	76%	53%	70%
T6-1	29	18	89%	0	75%	50%	66%
T6-2	13	6	100%	0	100%	83%	92%
T7-2	26	12	100%	0	92%	75%	89%
T9-1	29	18	83%	27%	87%	53%	72%
T9-2	22	12	92%	9%	100%	55%	77%
Median Scores	26	12.5	93.5%	4%	91.5%	54%	75%

Components evaluated

- Automatic Speech Recognizer (ASR)
- Interpreter
- ASR + Interpreter
- Dialogue Manager

Component Evaluation Metrics

- Compare system results with replicable human coding (Gold Standard)
- Basic Scoring Methods
  - Precision (correct recognized/ all recognized)
  - Recall (correct recognized / all correct)
  - F-Score (harmonic mean of P & R)
  - Error Rate (errors / all correct)
- Dialogue Measures
  - Over whole dialogue
  - Average of scores of each utterance in the dialogue

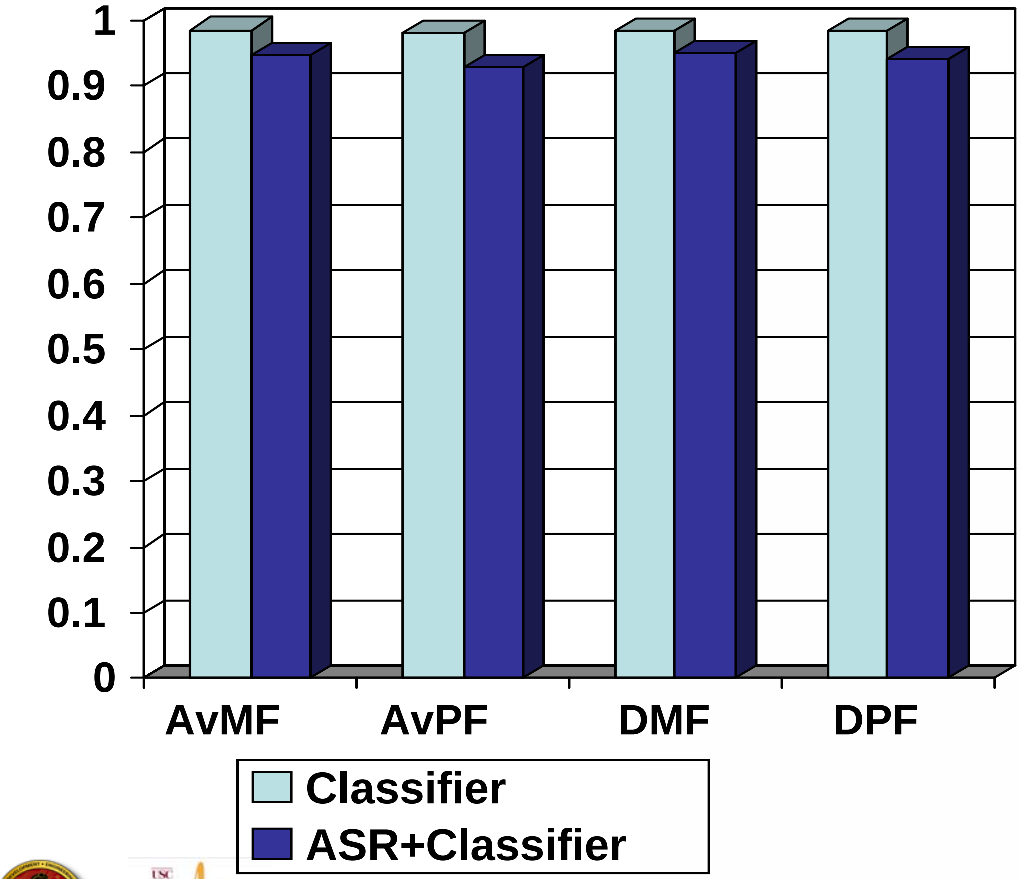
Example: ASR evaluation

- Transcribed Utterance (Exact reproduction of audio signal)
  steel one nine this is gator niner one adjust fire over
- Output from ASR
  steel one nine this is gator one niner one adjust fire over
- Merged view
  steel one nine this is gator [one] niner one adjust fire over
- Measures
- Precision = 11/12
- Recall = 11/11
- WER = 1/11
- F-Score(Harmonic mean of Precision and Recall) = 0.957

Evaluation Results: ASR scores

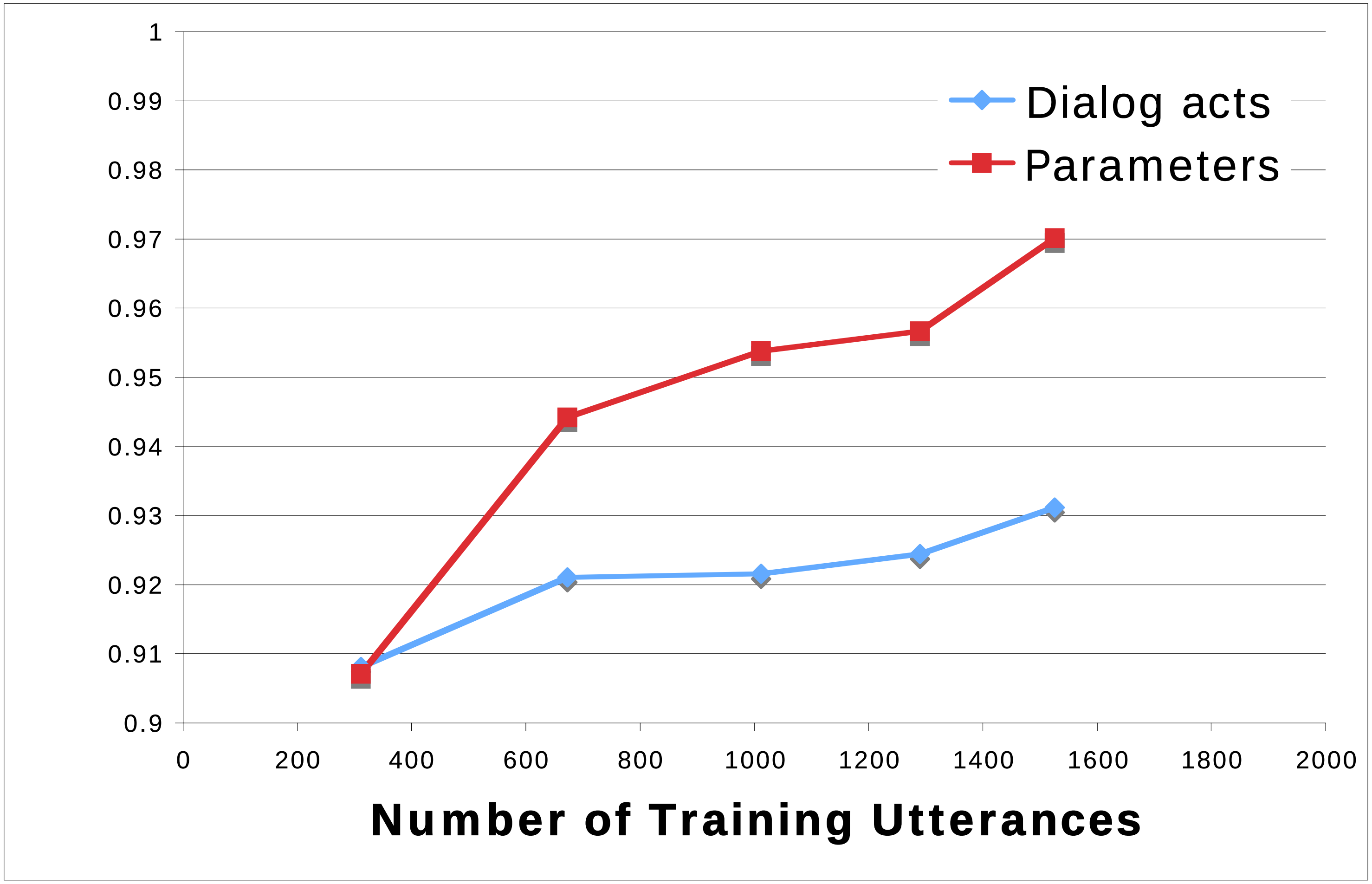
- Dialogue precision score (DP) = 0.900
- Dialogue recall score (DR) = 0.920
- Dialogue F score (DF) = 0.910
- Dialogue Word Error Rate (DWER) = 0.114
- The average precision score is (AvP) = 0.920
- The average recall score (AvR) = 0.935
- The average F score (AvF) = 0.927
- The average word error rate (AvWER) = 0.097

Interpreter vs ASR+Interpreter



- Interpreter Evaluation
  - Interpreter results on perfect input compared to human coding
- ASR + Interpreter Evaluation
  - Interpreter coding on ASR output compared to human coding

Radiobot Interpreter performance related to size of training data



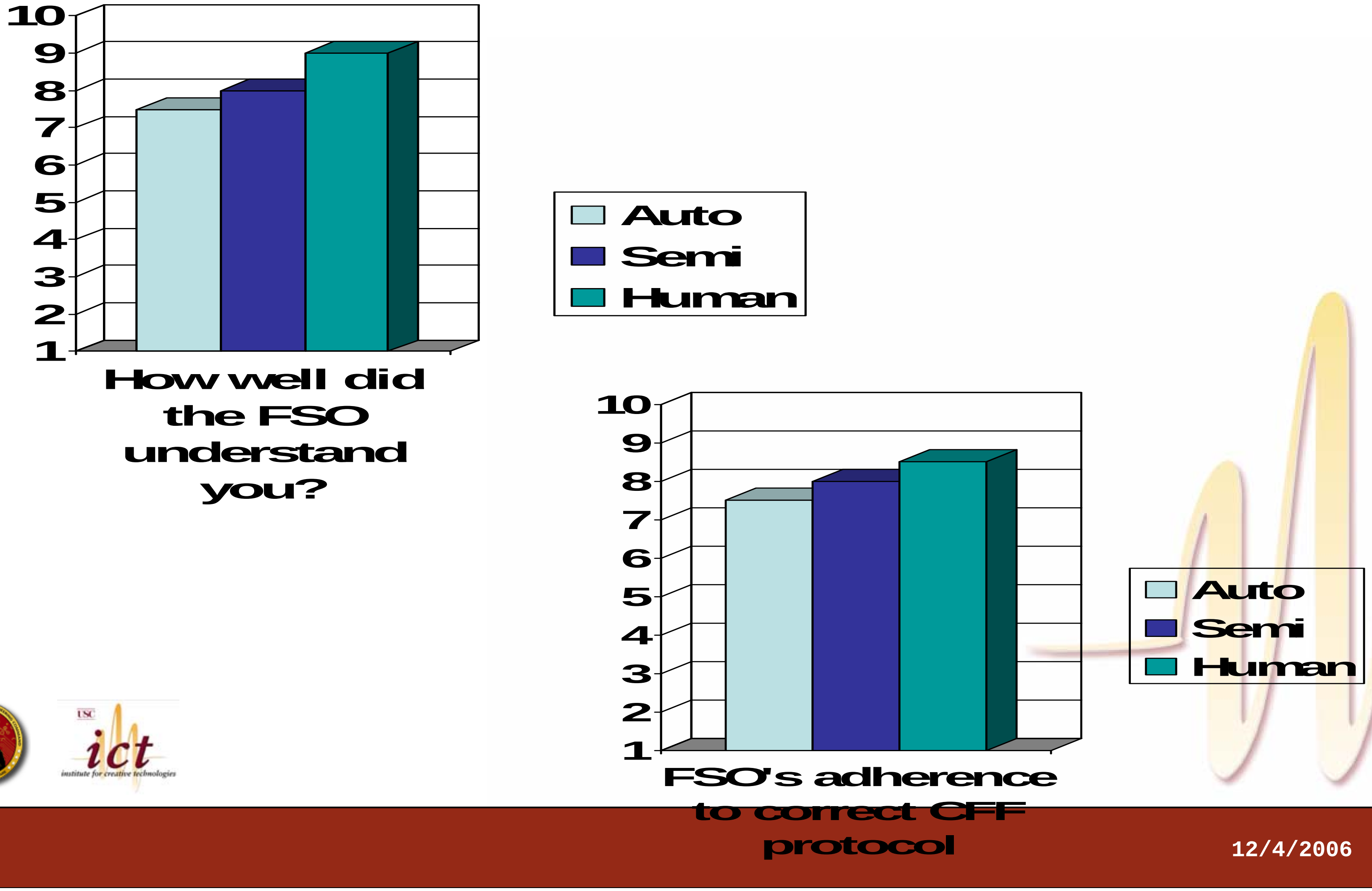
Dialogue Manager Evaluation

- Comparison of Machine coded Information state against human coded Information state.
- MACHINE:
  - has_warning_order true
    has_target_location false
    has_grid_location false
- HUMAN:
  - has_warning_order true
    has_target_location false
    has_grid_location false
- DIsER, DIsP, DIsR…, AvIsER, AvIsP…

Dialogue Manager scores

- Dialogue Information State Error Rate (DIsER) = 0.0106
- Dialogue Information State Precision (DIsP) = 0.9893
- Dialogue Information State Recall (DIsR) = 0.9893
- Dialogue Information State F score (DIsF) = 0.9892
- Average Information State Error Rate (AvIsER) = 0.0106
- Average Information State Precision (AvIsP) = 0.9893
- Average Information State Recall (AvIsR) = 0.9893
- Average Information State F Score (AvIsF) = 0.9893

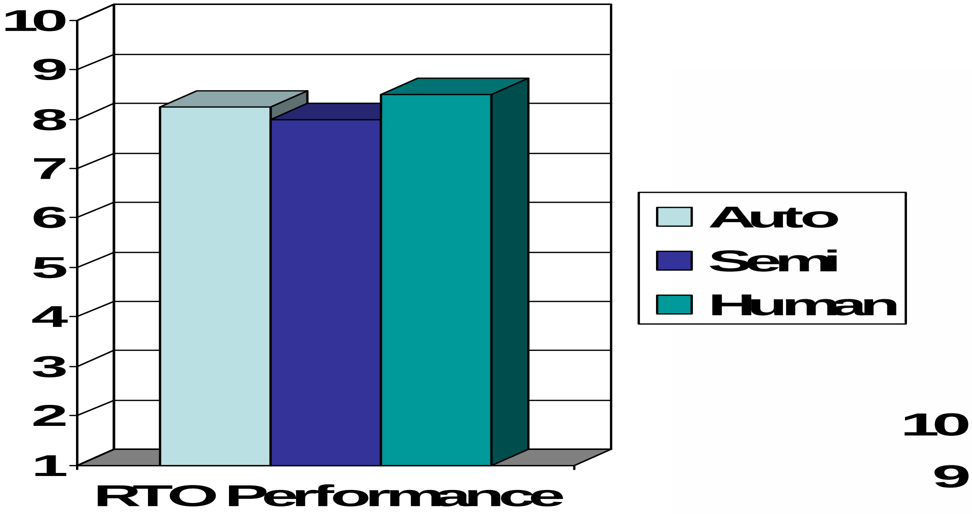
Questionnaire Results: Dialogue

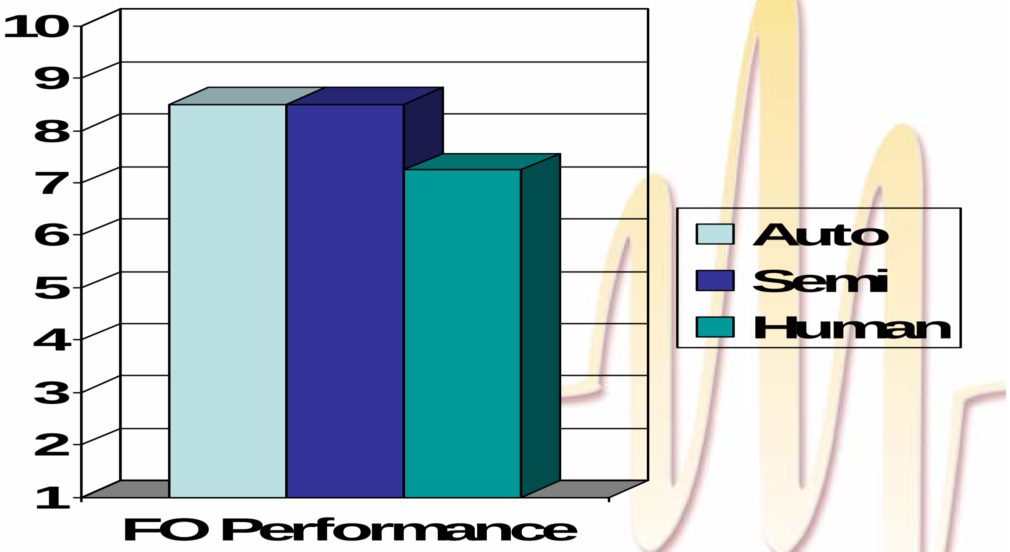


User Survey Feedback

- Near-human level quality on understandability and adherence to protocol
- Subjective judgments of trainee and partner (FO & RTO) performance higher or the same for Radiobot compared to human FSO

Questionnaire Results: Trainee Performance





Current Status

- Achievements
  - Allows large range of mission types (e.g., adjust fire, fire for effect, offset from known position, polar, grid)
  - Good performance on calls from men with standard American accent
- Needs work:
  - Improve recognition rate on Range of speakers (including female, regional accents, and non-native speakers (e.g. coalition forces)
  - Improve error handling due to recognition errors
  - Improve transparency and prompting
    - E.g. answer why firesim denies missions
  - Hardware robustness

Next Steps

1. Improving UTM Radiobots to performance level capability
   - Suitable for use in regular training
   - Improved error handling and feedback
   - Multiple synchronous missions
   - Better performance on wider range of speakers
   - multiple use cases, trainer aids, AAR aids
2. Adaptation to other CFF domains & platforms
   - Other parts of JFETS
   - Laptop trainer
   - Mobile/field use

Radiobot Future Plans

- Produce useful automation of radio communication in training simulations
  - off-load tasks from operator controller
  - standardize training
- Extension to other domains
  - E.g., 9-line, sitreps, fraternal unit communication
- Toolkits for non-expert radiobot construction for new domains

Soldiers with UTM Radiobot

QuickTime™ and a
Photo - JPEG decompressor
are needed to see this picture.